Dataset

Dataset yang digunakan didapatkan dari *repository* https://github.com/feryandi/Dataset-Artikel sebanyak 100 artikel dengan topik acak dan 20 artikel dengan topik staycation yang didapatkan dari berbagai portal berita di Indoneisa.

Pengujian

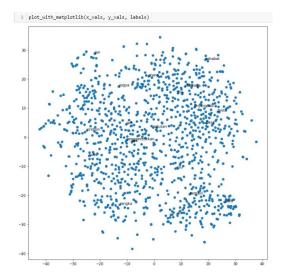
Perlu diketahui bahwa penulis melakuan 2 kali percobaan dengan pengaturan model yang berbeda, percobaan pertama penulis mengatur model dengan ukuran vektor = 200, negative = 10, iterasi epoch = 80 dan min_count = 1 sedangkan untuk percobaan kedua hanya menganggti min_count menjadi 5 dan sisanya sama seperti percobaan pertama. Penulis mengatur ukuran vektor 200 karena berdasarkan dokumentasi Word2vec ukuran yang disarankan berkisar puluhan hingga ratusan.

Hasil

A. Menggunakan Min count = 5

```
print(w2v_model.wv.similarity('staycation', 'liburan'))
print(w2v_model.wv.similarity('cirebon', 'hotel'))
print(w2v_model.wv.similarity('cirebon', 'kepala'))

0.6793084
0.1645842
-0.31514162
```

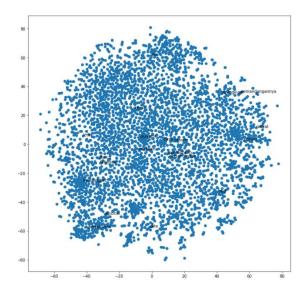


B. Menggunakan Min_count = 1

0.24269292 -0.27509636

```
print(w2v_model.wv.similarity('staycation', 'liburan'))
print(w2v_model.wv.similarity('cirebon', 'hotel'))
print(w2v_model.wv.similarity('cirebon', 'kepala'))

0.7806926
```



Analisa

Dapat dilihat pada tiap hasil percobaan ketika menggunakan min_count sebesar 5, kemiripan kata staycation dan liburan mendapatkan hasil yang lebih sedikit daripada min_count sebesar 1. Hal ini dikarenakan artikel pada topik staycation banyak membahas mengenai staycation yang merupakan alternative liburan. Sehingga kedua kata tersebut menjadi sangat mirip ketika min_count sebesar 1 dikarenakan banyakanya data pada artikel tersebut. Kemudian pada percobaan yang paling tidak mirip adalah pemilihan kata Cirebon dan kepala. Hal ini dapat kita asumsikan benar, dikarenakan Cirebon merupakan salah satu kota di Indonesia sedangkan kepala adalah bagian tubuh manusia, sehingga kata tersebut seharusnya tidak mirip dan dibenarkan dalam hasil akhir kedua model.

Kemudian dapat dilihat dari grafik plot kedua model memiliki hasil pola yang sama tetapi pada min_count 5 titik-titik katanya semakin sedikit dikarenakan 100 artikel dengan topik acak yang digunakan yang memungkinkan kata-kata tersebut tidak mencapai 5x keluaran, sedangkan ketika menggunakan seluruh kata (min_count = 1) seluruh kata diplot.